VAWKUM Transactions on Computer Sciences
http://vfast.org/journals/index.php/VTCS@ 2019, ISSN(e):2308-8168, ISSN(p): 2411-6335
Volume 16, Number 1, January-April, 2019 pp. 26–**30**

CLASSIFICATION OF TWITTER'S DATA TO GET GENDER IDENTIFICATION

WAQAS ALI[*1], MALIK TAHIR HASSAN[1], SYED FAWAD RAZA1, USMAN FIAZ1

[1]Department of Computer Science, University of Management and Technology, Lahore, Pakistan
Email: waqasali@umt.edu.pk

ABSTRACT. This paper describes the accuracy of various algorithms for classification of text on the basis of gender identification. We examined the knowledge extracted from corpus of twitter's online social media in term of gender identity. By comparing algorithms on different feature sets, we established a feature set of 20 distinct arguments which increase the correctness of gender identification on all over the twitter. We reported accuracies of three algorithms obtained by using two approaches applied on two classes of gender i.e. male and female; a model where a lot of features are reduced using powerset transformation.
Keywords*:* Stylometric feature . content based feature . Gender Identification . Data mining

1. Introduction. There are so applications in the field of data mining in which we get benefits from consistent models for detecting gender of users in social media. Textual posts such like tweets, facebook post or email show a large part of user created data, and consist of material which is related in determining the hidden user characteristics, or examining the honesty of self reported gender.

A very simple method for discovery of unknown user characteristics of social media applications is to knowledge the scripting ways of user by mining numerous characteristics from scripts they have shared or uploaded. This method, though, work best on the data of one social media application but it will not correctly work on the other social media websites or application at the same time.

This paper addressed the problems of reporting user of online social media application where a user can write something and then post, by choosing a characteristics set presented to have comparatively high predictive value in way of gender precision through numerous media types. In our proposed model, we have justify the gender identification opportunity of text to be classified as tweet. Our model marks an knowledgeable presumption about gender value of the user (female/male).

In addition, we observed the precisions got by two different approaches of feature extraction that are stylometric approach and content based approach, to the classification problem of instantaneously classifying gender. We applied both the feature sets extracted from the stylometric and content based approaches to three different algorithms wherein the feature set is reduced to a small set using powerset conversion.

The structure of paper is that in part 2 after reviewing related work, we described data set and numerous preprocessing stages in part 3. In parts 4 and 5, we gave a report of features and characteristics that we extracted and two approaches of feature extraction, reduction and data classification, the results are described in part 6.

2. Related work. In data mining, a lot of related work done in the filed of gender identification by using the textual data gathered from different social media applications, a large part of it is finished in reaction to the PAN 2013 user related information task [1]. Other recent work similar with this type of classification is done by Mowery, D., Bryan, C., & Conway in 2017, in which Classification of depression is done on data of twitter by collecting depressive features from it. The input data set is made up of 9300 twitter tweets which are based on hierarchical data model is given to the scikit-learn 0.18, SVM with linear kernel and five-fold cross-validation applied on the data which conclude 2644 evidence of depression with depressive symptoms 1656 (depressed mood-1010, depressed sleep-98 and fatigue-427) [2]. Other new work into user reporting has established the capability to conclude the unseen characteristics of user of social media with precisions of 91.5% for attribute like gender [3]. Works of this type, though, increase to emphasis on groups of extensive textual tweets. Similar effort has done on assuming inactive user features such like age, gender, political location and provincial source from most smaller social media messages and posts, as Netlog chat texts[1] and micro-blogs of Twitter [4]. In another research they have taken 50 participants whose expressiveness want to be judged. 25 among them were men and 25 were women. Age range was from 17 to 25 years and all were students. Then after there permissions there facebook and twitter accounts was examined for gain some expression markers to judge their expressiveness. After examining their accounts 6 techniques were

derived from their tweets, comment etc. Those 6 expressive Markers was: (1) Punctuation marks (e.g. !!) (2)Extensive Full Stops used with speech sample (e.g. ……) (3) Use text with capitalization (e.g. LAHORE) (4) Use of same letter within a word (e.g. nooooooo) (5) emoticons (e.g. :-) , >:O , :-O) (6) Words that express laughter (e.g. Hahaha, LOL). From all this record researcher has examined from this data of facebook and twitter women are more emotionally expressive as compare to men. From all these stereotypes researcher has examined that sadness, happiness and fear are believed to be the characteristics for women and men are characteristically more angry. Researcher has explored that examining a group of these fifty test members and their uses of expressive markers in online communication (social media) and got the hypothesis that social networking has influenced a lot of changes in communication between both genders. This paper conclude that women will still be the more emotionally expressive gender [5].

3. Data set and preprocessing. The data which we used to train our models is consist of data set downloaded from the tweeter's feeds and tweets. Two gender groups are involve in the target audience male and female. A corpus consist of total 425 accounts and each account has archive file of his/her overall tweets which he/she posted from the activation of account till now. Each archive is downloaded from tweeter official website in the form of text file and each text file consist of round about 600 to 700 tweets written in English language. Whole data is manually classified into two groups, and each group is labeled with tag of male and female respectively. In 425 groups, there were 215 groups were consist of female accounts and the other 210 accounts were consist of male accounts. The proportion of each gender is corpora is presented in Table 1.

Table 1: Proportion of each gender in corpora

Genre	English	
Twitter	Female	Male
	215	210

4. Feature extraction. To construct our feature set for gender identification we extract many different classes of features, getting inspiration from the relevant work such as features extracted from sentiment analysis [6], emojis used for emotions [7] and word count features. We catagories these features into two main catagories, first one is stylometric based features and the other one is content based features. In stylometric base feature detection, we manually read the files and construct a feature set of 20 elements which were frequently occurred in the files. The list of features is ('NoOfCommas' , ' OpenBracket' , ' FullStop' , ' Space' , 'QuestionMark' , ' NotSign' , ' LowerCase' , ' UnderScore' , ' Dash' , ' CloseBracket' , ' AtTheRate' , ' Digits' , ' AndSign' , ' SemiColons' , ' UpperCase' , ' Colons' , ' FarwordSlash' , ' HashSign' , ' EqualSign' , ' Percentage' , ' Gender' {' Female',' Male'}. Figure 1, is representing the values of these features in sample files.

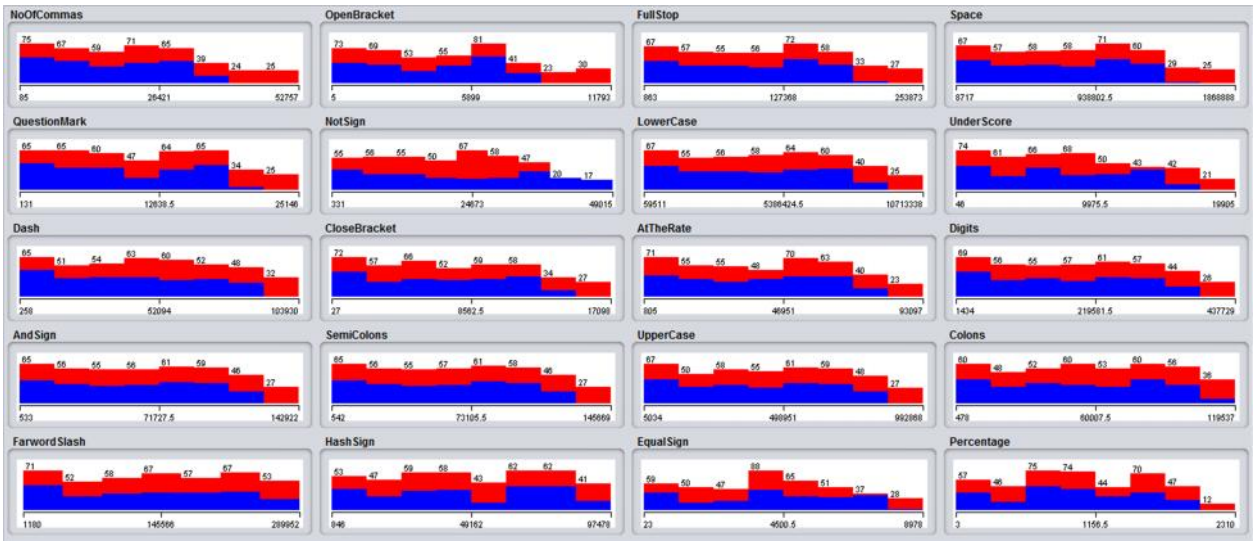

Figure 1: Representing the values of stylometric features

The other technique which we used to extract feature is content based method. In this method we use NGram and NCharacterGram Techniques. Weka is used to extract the feature set on the basis of these two techniques. Figure 2 shows the values of different grams used for feature detection:

Method Name	No of Features	Observation
word 1 Gram	1230	Every feature is single distinct word
Word 2 Gram	1346	Every feature is two distinct words
Word 3 Gram	1662	Every feature is three distinct words
Character 2 Gram	1072	Every feature is two distinct character
Character 3 Gram	1119	Every feature is three distinct character
Character 4 Gram	1184	Every feature is four distinct character
Character 5 Gram	1213	Every feature is five distinct character
Character 6 Gram	1280	Every feature is six distinct character
Character 7 Gram	1319	Every feature is seven distinct character
Character 8 Gram	1341	Every feature is eight distinct character
Character 9 Gram	1369	Every feature is nine distinct character

Figure 2: Representing the values of different feature extracted by using different Gram Values

These were very large amount of features and most of the features were useless and they were not frequent too. So we need to reduce the number of feature set. For this we used three different feature selection algorithms:

1. CfsSubsetEval (correlation based subset), it evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them.
2. Relief Attribute Eval (Relief), Evaluates the worth of an attribute by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same and different class.
3. InfoGainAttributeEval, Evaluates the worth of an attribute by measuring the information gain with respect to the class

Datasets Extracting Using Following Method	Feature Selection Method	No Of Feature after applying Selection Method
Word 1 Gram	CfsSubsetEvel(correlation based subset)	48 Reduces no of attributes
Word 2 Gram	ReliefFAttributeEval (Relief)	1346 ranks the attribute
Word 3 Gram	ReliefFAttributeEval (Relief)	1662 ranks the attribute
Character 2 Gram	InfoGainAttributeEval	1072 ranks the attribute
Character 3 Gram	InfoGainAttributeEval	1119 ranks the attribute
Character 4 Gram	InfoGainAttributeEval	1184 ranks the attribute
Character 5 Gram	CfsSubsetEvel(correlation based subset)	29 Reduces no of attributes
Character 6 Gram	CfsSubsetEvel(correlation based subset)	19 Reduces no of attributes
Character 7 Gram	ReliefFAttributeEval (Relief)	1319 ranks the attribute
Character 8 Gram	CfsSubsetEvel(correlation based subset)	31 Reduces no of attributes
Character 9 Gram	CfsSubsetEvel(correlation based subset)	30 Reduces no of attributes
Stylometry Based	InfoGainAttributeEval	1230 ranks the attribute

Figure 3: Representing the relected reduced number of features

5. Models and Evaluation. Two classification algorithms are used to train the dataset Randomforest and Jeccard coefficient (J48). RandomForest is a version of decision tree and work in the form of hierarchal structure. While J48 works on the dot product. Figure 4 shows the formula of J48. There were two datasets one that is got by applying feature extraction method and other dataset is that got bby applying feature selection method on same above dataset. Both the algorithms applied on the dataset and different results were computed and compared. Table 2 shows the result of word 1 Gram 2 Gram and 3 Gram, Table 3 shows the result of Character 2 Gram Character 3 Gram and Charater 4 Gram before and after feature selection, Table 4 shows the results of Character 5 Gram Character 6 Gram and Character 7 Gram before and after feature selection and Table 5 shows the results of Character 8 Gram and Character 9 Gram before and after feature selection with stylometry and content based features. Figure 4. Representing the gender identification model our proposed methodology.

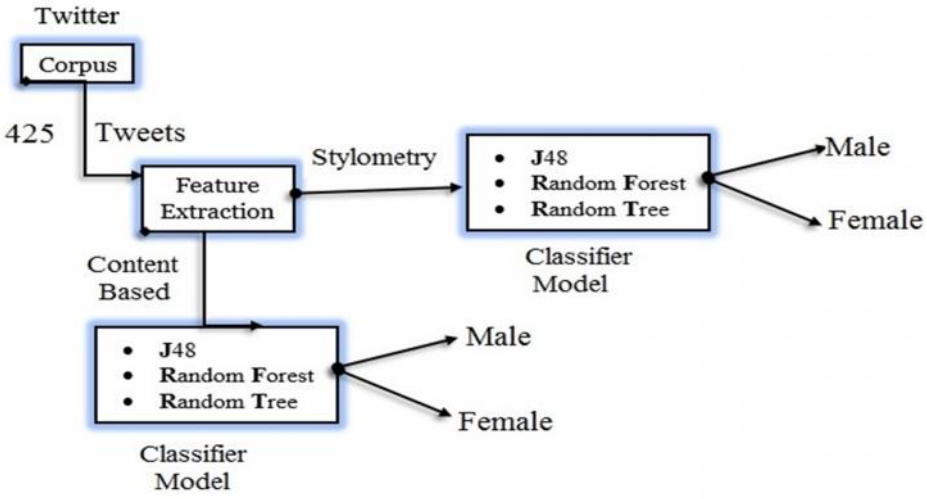

Figure 4: Representation of Gender Indentification Model

Table 2: shows the result of word 1 Gram 2 Gram and 3 Gram

Word	1Gram	2Gram	3Gram	1Gram After Feature Selection	2Gram After Feature Selection	3Gram After Feature Selection
Random Forest	70.99	67.72	65.58	71.39	68.4	66.13
J48	63.43	58.8	61.05	68.18	59.51	59.23

Table 3: shows the result of Character 2 Gram Character 3 Gram and Charater 4 Gram before and after feature selection

Character	2Gram	3Gram	4Gram	2Gram After Feature Selection	3Gram After Feature Selection	4Gram After Feature Selection
Random Forest	64.24	65.51	67.22	64.47	65.88	67.29
J48	55.46	58.33	59.46	56.28	58.19	59.55

Table 4: shows the results of Character 5 Gram Character 6 Gram and Character 7 Gram before and after feature selection

Character	5Gram	6Gram	7Gram	5Gram After Feature Selection	6Gram After Feature Selection	7Gram After Feature Selection
Random Forest	69.23	68.8	69.67	65.61	65.44	70.17
J48	60.06	57.34	59.38	63.31	66.16	59.5

Table 5: shows the results of Character 8 Gram and Character 9 Gram before and after feature selection with stylometry and content based features

Character	8Gram	9Gram	Stylometry Feature Based Dataset	8Gram After Feature Selection	9Gram After Feature Selection	Stylometry Feature Based Dataset after feature selection
Random Forest	69.76	68.99	98.8	69.34	65.44	98.83
J48	60.07	61.25	97.72	68.19	66.59	97.62

6. Discussion, Conclusion and Future Work. Analysing data we can say increase performance of j48 applying feature selection cfssubsetevel on character n gram dataset but decrease performance of random forest. But on word n gram applying feature selection CfsSubsetEvel increase performance of both Random Forest and j48. Applying relief feature selection method on dataset preformation of both classifier increases. Applying info gain feature selection method on dataset preformation of both classifier increases litilely. Random Forest performance is better than J48 classifier. Stylometry Feature based method is good for Gender Classification on our Corpus. Random Forest give 99% accuracy on dataset that is extracted using Stylometry based Feature Method. This is optimal solution.

Stylometry based feature sets give best Result according our observation. This work can be extended to get the more information about the writer/user of twitter by using more efficient algorithms to identify the age range of a specific user. And this work can be extended in the form to get investigation of the unethical activities done by user on social media that are involved in social crimes.

REFERENCES

[1]. Peersman, C., Daelemans, W., & Van Vaerenbergh, L. (2011, October). Predicting age and gender in online social networks. In Proceedings of the 3rd international workshop on Search and mining user-generated contents (pp. 37-44). ACM.

[2]. Mowery, D., Bryan, C., & Conway, M. (2017). Feature studies to inform the classification of depressive symptoms from Twitter data for population health. arXiv preprint arXiv:1701.08229.

[3]. Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one, 8(9), e73791.

[4]. Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010, October). Classifying latent user attributes in twitter. In Proceedings of the 2nd international workshop on Search and mining user-generated contents (pp. 37-44). ACM.

[5]. Parkins, R. (2012). Gender and emotional expressiveness: An analysis of prosodic features in emotional expression. Pragmatics and Intercultural Communication, 5(1), 46-54.

[6]. Mukherjee, A., & Liu, B. (2010, October). Improving gender classification of blog authors. In Proceedings of the 2010 conference on Empirical Methods in natural Language Processing (pp. 207-217). Association for Computational Linguistics.

[7]. Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010, October). Classifying latent user attributes in twitter. In Proceedings of the 2nd international workshop on Search and mining user-generated contents (pp. 37-44). ACM.